DOCUMENT RESUME

ED 357 041 TM 019 751

AUTHOR Kim, Haeok; Plake, Barbara S.
TITLE Monte Carlo Simulation Comparison of Two-Stage Testing and Computerized Adaptive Testing.
PUB DATE Apr 93
NOTE 43p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Ability; *Adaptive Testing; Comparative Testing; *Computer Assisted Testing; Computer Simulation; Estimation (Mathematics); Individual Differences; Item Response Theory; *Monte Carlo Methods; Test Items; Test Length
IDENTIFIERS Ability Estimates; *Routing Tests; *Two Stage Testing

ABSTRACT

A two-stage testing strategy is one method of adapting the difficulty of a test to an individual's ability level in an effort to achieve more precise measurement. A routing test provides an initial estimate of ability level, and a second-stage measurement test then evaluates the examinee further. The measurement accuracy and efficiency of item response theory (IRT) based two-stage testing was investigated in comparison with an individualized computerized adaptive test (CAT). Eighteen simulated two-stage tests and three fixed-length CATs differing in the number of test items administered were compared. Abilities were generated for a sample of 1,600 simulees. Results indicate that the statistical characteristics of the routing test have a major influence on measurement precision in ability estimation. Overall, it was apparent that a fixed-length CAT is superior to the two-stage tests of equivalent length in terms of measurement accuracy and efficiency. IRT-based two-stage tests using rectangular distribution of item difficulties in the routing test and an odd number of second-stage tests produced more accurate theta estimates than did the other two-stage test configurations studied. IRT-based two-stage tests may sometimes be practical alternatives to CAT, considering its limitations. Two tables and 18 graphs present analysis results. (SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
HAEOK KIM
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED357041

# Monte Carlo Simulation Comparison of Two-Stage Testing and Computerized Adaptive Testing

Haeok Kim

Barbara S. Plake

University of Nebraska-Lincoln

A paper presented at the annual meeting of the National Council on Measurement in Education, Atlanta, April, 1993.

TM019751



BEST COPY AVAILABLE

# INTRODUCTION

Fixed-length paper-and-pencil conventional tests can assess large numbers of individuals quickly and inexpensively. With conventional tests, all examinees take the same test irrespective of their latent ability on the construct being measured. In contrast, an "adaptive test" presents different test items to different individuals as a function of each individual's estimated status on the trait being measured.

A two-stage testing strategy is one method of adapting the difficulty of a test to an individual's ability level in an effort to achieve more precise measurement. Two-stage testing involves an initial routing test followed by second-stage measurement test. The routing test provides an initial estimate of an individual's ability level. Based on this score, the examinee is then assigned to one of several second-stage "measurement" tests, chosen as a function of the examinee's estimated ability from the routing test. Ability estimates for the examinees are then derived by combining their scores from the routing test and the second-stage measurement test (Lord, 1971, 1980).

Two-stage testing has been examined by several research studies focusing on their comparison to conventional tests (Angoff and Huddleston, 1958; Cleary, Linn, and Rock, 1968; Linn, Rock, and Cleary, 1969; Lord, 1971, 1980; Loyd, 1984; Betz and Weiss, 1973, 1974). Results are generally favorable for the two-stage tests, showing a reduction in test length without degrading measurement

accuracy. However, these applications of two-stage tests in early research were not based on the Item Response Theory (IRT) and therefore had several limitations (Weiss, 1982). A primary limitation was that most of earlier studies of two-stage testing generally used only item difficulty information to structure the item pool. Item selections, therefore, did not make full use of item information since the algorithm employed ignored item parameters of discrimination and guessing susceptibility. Further, scoring methods used in these non-IRT based two-stage tests were not appropriate when different items are answered by different examinees. These limitations of two-stage tests can be circumvented through Computerized Adaptive Testing (CAT), an IRT approach to adaptive testing. In CAT, each examinee's ability level is iteratively estimated during the testing process. Typically, CAT first administers an initial item, often from the middle of the prospective ability range. If that question is answered correctly, the next question administered is usually more difficult. If the question is incorrectly answered, the next one is typically easier. Often this item administration decision process continues until the examinee's ability is measured to prespecified degree of accuracy.

A general finding of CAT research is these test scores have reliabilities and validities equal to or greater than the those of scores from comparable conventional tests (Weiss, 1982), even with reductions in test lengths up to 50% (Olsen, Maynes, Salwson, & Ho, 1986). In addition, CAT typically requires less test

administration time to reach an equivalent level of precision (Moreno et al., 1984).

Although there is no doubt that CAT provides an efficient and accurate ability score estimates, there are technical and practical constraints to the implementation of CAT. These limitations include concerns about content balance, possibility of context effects, hardware limitation, and cost. Some of the problems arising from context effects and content validity can be reduced through a two-stage test approach since the routing and second-stage tests are constructed prior to administration. Recently, Adema (1990) proposed mixed integer linear programming (MILP) models, containing continuous and integer decision variables, for the construction of paper-and-pencil two-stage tests. These models take into account practical constraints which can control test composition, administration time, inter-item dependencies, and other practical problems encountered with CAT.

There are some additional negative effects possible from taking a CAT. In typical CAT, the opportunity of reviewing and altering responses generally has not been allowed. Two-stage tests with paper-and-pencil administration are relatively free from these limitations associated with CAT, including the loss of control by examinee to skip, change, or review items, and the high cost of incorrectly entered answers.

Two-stage tests have the advantage that they can be administered by paper-and-pencil. In comparison with conventional tests, two-stage tests are potentially more accurate in situations

where the group tested has a range of ability too wide to be measured effectively by a typical conventional test (Lord, 1980).

The purpose of this study was to investigate the measurement accuracy and efficiency of IRT-based two-stage testing in comparison to an individualized CAT and to ascertain the conditions when two-stage test might be an acceptably close alternative to CAT in terms of accuracy of measurement. For two-stage tests, different combinations of lengths of the routing test (10, 15, and 20), distributions of item difficulty parameters in the routing tests (peaked and rectangular), and differing number of second-stage tests (6, 7, and 8) were simulated. A total of eighteen two-stage tests and three fixed-length CATs, differing from one another in the number of test item administered as a terminating criteria (40, 45, and 50), were compared to address following research questions:

1) What is the relationship among ability estimates derived from the two-stage testing, those from CAT, or hypothetical underlying ability?
2) What is the accuracy of ability estimates obtained from the two-stage testing and CAT in comparison to hypothetical underlying ability?
3) What is the amount of information or precision provided by each testing strategy at various points along the ability continuum?
4) What is the relative efficiency of the varying two-stage tests in comparison to the CAT?

## METHODOLOGY

For the first stage of the simulation study, abilities for a sample of 1600 simulees were generated by creating 100 thetas at each of 16 discrete ability levels at and between -3.0 and 3.0 (i.e., -3.0, -2.6, -2.2, -1.8, -1.4, -1.0, -0.6, -0.2, 0.2, 0.6, 1.0, 1.4, 1.8, 2.2, 2.6, 3.0). A separate item pool for the CAT and two-stage tests was generated similar to each other. Each item pool consisted of 354 items with 5-alternative responses and had the discrimination parameters (*a*) and the pseudo-guessing parameters (*c*) fixed at values of .7 and .18, respectively. For the use with CAT administration, the item difficulty parameters (*b*) were uniformly distributed in the range of -3.0 and 3.0, with at least four items at each 0.1 interval. For use with the two-stage tests, the difficulty parameters (b) were generated to satisfy special characteristic of the routing tests and second-stage tests. Detailed information is addressed in the test construction section. The modified one-parameter logistic IRT model was used to simulate all item responses and to select items.

### Construction of Two-stage test

For the two-stage tests, the items for the routing tests and the second-stage tests were selected from the same item pool. However, the items selected for the routing tests were not used again for the second-stage test.

**Routing Tests:** For the routing test with a peaked distribution, the items were peaked at the median difficulty level (item difficulty range of -0.4 to 0.1). Items were varied from very easy

items to very difficult items, with a difficulty range of -3.0 to 3.0, for the rectangularly distributed routing test.

**Second-Stage Measurement Tests:** Differing number of second-stage tests (6, 7, or 8) were developed, each composed of 30 items. Each measurement test was not as peaked as the routing test as indicating by the larger ranges and standard deviations of item difficulties for the routing tests. For economy of items and potential reduction for routing error, each second-stage measurement test included 60% overlapped items with adjacent levels of difficulty (30% lower, 30% higher) except at the lowest and highest level where there was only 30% overlap. The range of item difficulty for each second-stage measurement test was equally divided by the number of levels (6, 7, or 8) at and between 3.0 and -3.0. For example, for the design with 6 second-stage tests, the ranges of b values for the lowest through the highest level of the second-stage tests are -3.0 to -1.667, -2.067 to -0.733, -1.133 to 0.210, -0.199 to 1.133, 0.733 to 2.067, 1.667 to 3.0, respectively.

Administration and Scoring of Two-Stage Tests

Computer simulation of two-stage tests was carried out as follows: First, one of the 18 two-stage tests was selected:

a) characteristics of the routing tests (peaked or rectangular distribution of item difficulties)
b) length of the routing tests (10, 15, or 20 items)
c) the number of second-stage tests (6, 7, or 8)

Second, the two-stage test selected from 18 two-stage tests was administered to 100 simulees from each of the 16 ability levels.

Third, response vectors for the chosen two-stage test were used to obtain 100 ability estimates at each of the 16 ability levels. Ability estimates of the routing test were first computed by maximum likelihood estimation procedure using Bayesian priors. Based on the ability estimates of the routing test, simulees were assigned to one of alternative second-stage measurement tests which average difficulty (b) was closest to their $\Theta$ estimate from the routing test. For total ability estimates for the two-stage test (routing test and second-stage test), maximum likelihood estimation of $\Theta$ for 40-, 45-, 50-item two-stage tests were derived from combined item vectors from the routing test and the second-stage test.

Administration and Scoring of CAT

The CAT simulations were performed using the MicroCAT™ testing service program (MicroCAT; Assessment Systems Corporation, 1984). The CATs were administered to 100 simulees from each of the 16 levels of theta. The initial ability estimate for each simulee was set equal to -0.2, which was about in the middle of the difficulty range of the item pool. A maximum information item selection procedure was used and also the maximum likelihood scoring procedure was used to obtain ability estimates after each item is administered. In order to make CATs comparable to two-stage tests in terms of test length, ability estimates were calculated after 40, 45, or 50 items administered. After 50 items administered, CAT administration was stopped.

A total of 18 two-stage tests and three fixed-length CATs were

constructed. To make direct comparisons of two-stage tests and CATs, this study used item pools of equivalent size, the same underlying statistical model, the same maximum likelihood scoring procedure, and tests of equivalent length. For each fixed total test length (40, 45, or 50), 6 variations of two-stage testing (routing test: peaked or rectangular by number of second-stage tests: 6, 7, or 8) were compared to CAT in terms of accuracy of $\Theta$ estimation.

Data Analysis

Pearson product-moment correlation coefficients were calculated to investigate the relationship among ability estimates derived from the eighteen two-stage tests, those from the CAT, and the true theta values.

The second step in the analysis was to compare the ability estimates obtained from the two-stage tests and the CAT to the true trait level for the simulees. For each testing condition, the root mean squared error ($RMSE = (\Sigma(\hat{\Theta} - \Theta)^2/N)^{1/2}$) of ability estimates were calculated by computing the square root of the mean squared difference between true ability and estimated ability for each simulee at the 16 ability levels. In addition to RMSE, a bias analysis ($BIAS = \Sigma(\hat{\Theta}-\Theta)/N$) was conducted for two purposes: (1) to identify the extent of the bias in the maximum likelihood ability estimates and (2) to indicate whether the errors reflect a systematic tendency to overestimate or underestimate the ability. Since maximum likelihood ability estimates tend to be biased for finite test lengths, it is useful to investigate whether these

testing procedures show different levels of bias observed.

The third step in the analysis was to compare the observed test information functions from each testing condition. These information functions were examined to determine the loss of efficiency in two-stage testing under the conditions of distribution and length of routing tests and number of level of measurement tests as compared to the CAT.

## RESULTS

### Correlations

Based on the ability estimates from maximum likelihood scoring procedure, as shown in Table 1, the ranges of the correlations between theta estimates and their true abilities were 0.971 to 0.982 for the 9 two-stage tests with the peaked routing tests, 0.975 to 0.982 for the 9 two-stage tests with the rectangular routing tests. These degrees of association were smaller than those between the ability estimates from CATs and their true abilities ($r_{CAT40}$=0.983, $r_{CAT45}$=0.985, $r_{CAT50}$=0.987). The effect of statistical characteristic of the routing test was not noticeable. However, with 7 second-stage tests, the rectangular routing test represented higher correlation than did the peaked routing test across the test lengths. Further, of all the two-stage test simulations, the two-stage test combining of a 10-item peaked routing test with 7 second-stage tests had the lowest degree of association between estimated thetas and their true abilities (r=0.971). In general, the pattern of correlation between ability

estimates from the two-stage tests and those from the CAT was the same as the pattern of correlation from ability estimates from the two-stage tests and true abilities, but with slightly lower value of correlations (r=0.957 to 0.969).

RMSE

Average RMSE of ability estimates for the 40-, 45-, 50-item tests are presented on Table 2. Except for the two-stage tests with 7 second-stage tests, there was no systematic effect of the statistical characteristic of the routing test across the test length. When two-stage tests are combined with 7 second-stage tests, the two-stage tests with rectangular routing tests showed smaller average RMSEs than did the two-stage tests with peaked routing tests across the different test lengths (see Figure 1).

**Effects of test length:** In this study, the effect of test length reflects the effect of routing test length since each of the second-stage tests has the same length of 30 items. As shown in Figure 1.1, two-stage tests with longer routing test, in general, produced more accurate $\Theta$ estimates, as indicated by smaller average RMSEs.

**Effects of differing number of second-stage tests:** When the effect of differing number of second-stage tests was considered, mixed results were found depending on the routing test length. For the two-stage tests using a 10-item routing test, the two-stage tests with largest number of second-stage tests produced more accurate theta estimates, as shown by smaller mean RMSE, for each of statistical characteristic of the routing test. For example, for

the two-stage tests with a 10 peaked routing items (n=40), average RMSE for the two-stage test with 8 second-stage tests (m=0.402, sd=0.075) was smaller than those for the two-stage tests with 7 second-stage tests (m=0.420, sd=0.116) and 6 second-stage tests (m=0.404, sd=0.074). This pattern of the results is supported by the two-stage tests with the rectangular routing items, too. However, the differences in average RMSE between the 6 and 8 number of second-stage tests were quite small for each of the routing tests (see Table 2). With longer routing tests (15- or 20-item), the two-stage tests with 6 second-stage tests (n=45, 50) showed the lowest RMSE for the peaked routing tests and the most accurate $\Theta$ estimates were obtained using 7 second-stage tests for the rectangular routing tests. Further, for the rectangular routing tests, there was no difference between average RMSEs from the two-stage tests using 6 second-stage tests and those using 8 second-stage tests across the test lengths (n=40, 45, 50). Thus, with the 15 or 20 routing items, the effect of the increase in a number of second-stage test was not positive in terms of measurement accuracy. Graphic representation of these trends for each of statistical distribution of routing tests were shown on Figures 1.2 and 1.3, respectively.

**Effect of the odd (7) number of second-stage tests**: For the effect of the odd number of second-stage test design, the results of the RMSE from the two-stage tests with peaked routing tests were unexpected. It was anticipated that using the odd number (7) of second-stage tests would result the better theta estimates than

from an even number (6 or 8) of second-stage tests, to be indicated by smaller average RMSE. Although there was no considerable difference among average RMSEs (range of 0.018 to 0.001) from 6, 7, or 8 second-stage test designs, the two-stage test with 7 second-stage tests was least accurate in ability estimation across the test lengths. This result was mainly due to significantly larger RMSEs for extremely low theta ($\Theta$=-3.0) across the test lengths. Refer to Figures 2.1, 2.2, and 2.3 for graphic representation of these trend. Thus, using the odd number (7) of second-stage tests had a rather negative effect in accurate theta estimation for the two-stage tests with the peaked routing tests.

However, for the rectangular routing tests, the two-stage tests with 7 second-stage tests (n=45, 50) were most accurate as shown by the lowest average RMSE.

**Effect of distribution of routing test:** As shown in Figures 2.1, 2.2, and 2.3, the two-stage tests with the peaked routing tests had considerably smaller RMSEs for the approximate theta range of -0.6 to 1.0 than did two-stage tests with rectangular routing tests and some of these RMSE were even smaller than those of CAT. On the other hand, the two-stage tests with the rectangular routing tests had smaller RMSEs at extreme thetas ($\Theta \leq$-2.2 and $\Theta \geq$2.6) than did the tests with the peaked routing tests and the magnitude of RMSEs for the two-stage tests with the rectangular routing items was relatively constant across the 16 ability levels (Figures 3.1, 3.2, 3.3).

**Comparison of two-stage tests and CAT results:** Not surprisingly,

all three CATs had smaller average RMSEs than did any of the two-stage tests and longer CAT showed lower average RMSE than did shorter CATs ($m_{cat40}$=0.350, sd=0.026; $m_{cat45}$=0.322, sd=0.023; $m_{cat50}$=0.314, sd=0.031, respectively). Even though slightly larger RMSEs were reported at the extreme low thetas, the degree of RMSEs of the CATs was very constant across the 16 ability levels.

In the comparisons of the six 40-item two-stage tests (10-item routing test) and the 40-item CAT, the two-stage test combining with a peaked routing test and 8 second-stage tests had the closest value of average RMSE (m=0.402, sd=0.075) to average RMSE from the 40-item CAT (m=0.350, sd=0.026). Among the six 45-item two-stage tests (15-item routing test), the most accurate $\theta$ estimates were obtained using a rectangular routing test with 7 second-stage tests, where its average RMSE (m=0.364, sd=0.026) is comparable to that of the 45-item CAT (m=0.322, sd=0.023). For the two-stage tests with the 20-item routing test (50-item two-stage tests), again the two-stage test with a combination of rectangular routing items and 7 second-stage tests had the smallest average RMSE (m=0.356, sd=0.02). This value was close to that of 50-item CAT (m=0.314, sd=0.031).

Overall, the 50-item two-stage test combining the 20 routing items from a rectangular distribution and 7 second-stage tests showed the smallest average RMSE (m=0.356, SD=0.022) among the 18 two-stage tests. This magnitude of average RMSE was very close to average RMSE of CAT40 (m=0.350, SD=0.026) and comparable to those of CAT45 (m=0.322, SD=0.023) and CAT50 (m=0.314, SD=0.031). This

result is presented graphically in Figure 1.

Bias Analysis

The bias index does not indicate the degree of estimation accuracy in an absolute sense because equal positive and negative errors would result in a zero bias by canceling each other. Rather, bias index is an indicator of whether there is a systematic tendency to overestimate or underestimate the ability parameter. In general, it was noticed that the number of positive bias values (overestimation of true ability) was more than the number of negative bias values in the two-stage tests and the CATs. Refer Figures 4.1, 4.2, 4.3, 5.1, 5.2, and 5.3 for graphic representation of these results.

Although all of the 18 two-stage tests reported bias of ability estimates across the 16 ability levels, irrespective of the test length (n=40, 45, 50), the two-stage tests combining a peaked distribution of item difficulties in the routing test and 7 second-stage tests showed less accuracy by underestimating the lowest theta ($\Theta$ =-3.0), than did all the other two-stage tests (see Figures 4.1, 4.2, and 4.3).

Test Information and Relative Efficiency.

All of the 18 two-stage tests, irrespective of statistical characteristic of the routing tests and test length, were least informative at the extreme low end of the $\Theta$ scale ($\Theta$=-3.0). In general, for the CATs, there was a tendency to obtain more information for extremely high ability and significantly less information for extreme low ability. For the two-stage tests with

the rectangular routing items, the test information functions were relatively constant and, in general, substantially lower than those of the two-stage tests with the peaked routing items for the approximate theta range of -1.0 to 1.4 which overlaps with the item difficulty range of peaked routing items (-0.4 to 0.1). Except for theta values in a range from -1.0 to 1.4, the two-stage tests with rectangular routing items displayed more test information than those of the two-stage tests with the peaked routing items across the test length (n=40, 45, 50). For an approximate ability range of -0.6 to 0.6, the two-stage tests with the peaked routing items produced even higher test information than did their counterparts in CATs. For theta values outside this range, CATs yielded constantly higher information across the ability levels than two-stage tests with the peaked items.

The relative efficiency of the two-stage tests and CAT is determined by the ratio of the information functions, which can be interpreted as the increase in the test length of the test with lower levels of information required for it to measure at the same level of information as the more informative test (Lord, 1980). For these comparisons, the relative efficiency of test information function for each of two-stage tests (n=40, 45, 50) and CATs are plotted as shown in Figures 6.1, 6.2, 6.3, respectively. For example, as can be seen from Figure 6.1, the two-stage test with a 10-item peaked routing test measured with approximately the same level of information as the CAT with 40-item for the ability range of -0.6 to 0.6. By contrast, the two-stage test with a 10-item

rectangular routing test (n=40) would need 44 to 49 items to measure as well as the CAT with 40-item for these ability levels. At $\Theta$=-3.0, with 6 second-stage tests design, the average information of the two-stage test with the 10-item peaked routing test and the two-stage test with the 10-item rectangular routing test were 5.35 and 5.84, respectively, and that for the $CAT_{40}$ was 7.76; the two-stage test with the 10-item peaked routing test would need to be lengthened from 40 to 58 items and the two-stage test with the 10-item rectangular routing test would need to be lengthened from 40 to 54 items to measure as well as the 40-item CAT. Finally, at $\Theta$=3.0 the two-stage tests with 10-item peaked routing test would need 55 items to measure as well as the 40-item CAT, while the two-stage test with the 10-item rectangular routing test would require 51 items. As longer two-stage tests (n=45, 50) were compared to longer CAT (CAT45, CAT50), the two-stage tests were relatively less efficient across the ability level than were shorter two-stage tests.

## DISCUSSION AND CONCLUSIONS

The purpose of this study was to compare ability estimates from the eighteen two-stage tests, varying the test length (n=10, 15, 20) of routing tests, statistical characteristics (peaked or rectangular distribution of item difficulties) in the routing tests, and the number of second-stage tests (6, 7, or 8), and three CATs using a fixed-length stopping rule (n=40, 45, 50) and then identify under what conditions two-stage test might be an

acceptably close alternative to CAT in terms of accuracy of measurement.

Statistical Characteristics of Routing Tests.

The main results of the study indicate that the statistical characteristic of the routing test in the two-stage tests has major influence on measurement precision in ability estimation. As expected, administering the items spanning entire difficulty range of the item pool produce better ability estimates at the lower and higher ability levels and administering the only items around the median difficulty of the item pool provides with better ability estimates at the middle ability levels. Therefore, these trade-off effects resulted in no considerable differences between average RMSEs from two different statistical characteristics of the routing tests. However, using 7 second-stage tests, the two-stage tests with the rectangular routing test are superior to the two-stage tests with the peaked routing test by showing smaller average RMSE across the test lengths. The two-stage tests combining 7 second-stage tests with the peaked routing test produced considerably less accurate ability estimates at $\Theta$=-3.0, indicated by larger negative bias index. Since the peaked routing items are in the range of b values -0.4 to 0.1, these routing items would be too difficult for extreme low ability examinees. However, with non-zero guessing parameter the examinees for low ability would get higher ability estimates than their true $\Theta$ values. Based on the overestimated abilities from the routing tests, examinees would be assigned to higher level of second-stage test, containing items too

hard for their abilities. As a result, they would obtain more wrong answers and end up with an underestimation of their true abilities. Thus, this bias trend might be explained by looking at the proportion of ability estimates that fell outside the restricted $\Theta$ range of -3 to +3. For the two-stage tests combining 7 second-stage tests with the 10 peaked routing items, at $\Theta$=-3.0 the proportion of ability estimates outside the restricted range ($\hat{\Theta}$<-3.0) was 52 out of 100 simulees. 25 simulees out of these 52 were assigned to higher level test ($m_b$=-1.59, sd=0.37) which is much harder than they should to be assigned ($m_b$=-2.45, sd=0.32) and resulted in larger negative bias. As noted by Hulin et al.(1982), the inclusion of a large number of extreme estimates of theta tends to distort RMSEs. Also, the amount of bias would be less if extreme low ability estimates are eliminated. For example, when the extreme theta estimates ($\hat{\Theta}$≤-4.0) were excluded (10 out of 52 simulees), the degree of RMSE was reduced to 0.492 from 0.814 and amount of bias was changed to 0.037 from -0.150 at $\Theta$=-3.0.

Test Length of Routing Tests

In addition to the statistical characteristic of the routing test, increasing the length of the routing test was important in reducing the size of the ability estimation errors. From the results presented, when the longer routing test is used, more accurate $\Theta$ estimates are obtained. However, if the routing test is too long, it will lose its efficiency in estimating examinees' abilities quickly. The results from this study suggest that a routing test length of 20 item was more desirable than routing test

lengths of 10 and 15.

Differing Number of Second-Stage Tests

In studying the effects of the number of second-stage tests, the increase in a number of second-stage tests from 6 to 8 does not make any improvement for the two-stage tests with the rectangular routing tests in terms of measurement accuracy and it was even less accurate for the two-stage tests with the peaked routing tests. For the two-stage tests with the rectangular routing items, it was clear that the two-stage tests using the odd number of second-stage tests yield better ability estimates than does those with the even number of second-stage tests. That is, the two-stage tests using 7 second-stage tests were superior to those with 6 second-stage tests and even those with 8 second-stage tests in terms of measurement accuracy. However, the effect of using the odd number of second-stage tests was not desirable with the peaked routing test since the two-stage tests combining the peaked routing tests and 7 second-stage tests produced larger average RMSE than did two-stage tests with even number of second-stage tests design across the test lengths. It was expected that using an odd number of second-stage tests would likely yield two-stage test scores that are more precise at the mean ability level. Actually, at around mean ability level ($\Theta$=-0.2, 0.2) the two-stage tests with 7 second-stage tests yielded more accurate ability estimates across the test length than did those with 6 or 8 second-stage tests. However, at $\Theta$=-3.0 these tests yielded considerably less accurate theta estimates than those with 6 or 8 second-stage tests. Thus, average

RMSE of the two-stage tests combining peaked routing tests and 7 second-stage tests were larger than those with an even number (6 or 8) second-stage tests. The difference in average RMSEs between the two-stage tests with 7 second-stage tests and the two-stage tests with even number of second-stage tests were reduced with longer routing tests.

If second-stage tests had not been designed to overlap at difficulty level, using an even number (6 or 8) of second-stage tests (half of second-stage tests at difficulty levels above the mean and half of second-stage tests at difficulty levels below the mean) would necessitate routing examinees of mean ability level up or down into a less appropriate second-stage test. However, this problem was addressed in this study by the overlap design of the second-stage tests. Under these conditions, there appeared to be no merit to use the odd number design for the two-stage tests with the peaked routing tests. However, with rectangular routing tests where items were spread throughout the theta range of -3 to +3, an odd number of second-stage tests was desirable, except for the shortest routing test (10-item). This result suggests that further studies on the effect of an odd number of second-stage tests need to take into account the contributing effects of an odd number of second-stage tests and an overlap design.

Overall, it is apparent from the comparison of the 18 two-stage tests and the three fixed-length CATs with respect to RMSE, bias, Pearson product-moment correlation coefficients, and test information, that the two-stage test with a combination of the 20

rectangular routing items and the 7 second-stage test produced ability estimates (RMSE= 0.356, sd=0.022) that may be considered as accurate as those of CAT with 40 items and comparable to the CATs with 45 items and 50 items. Although there is no absolute sense of interpretation for RMSE, this result (RMSE of 0.356 and a correlation of 0.982) exceeds Hulin et al's (1982) interpretation for an evidence of precise theta estimation (e.g., RMSE of 0.377 or less and a correlation of 0.927 or higher). The relative test efficiency determined by the amount of information from the two-stage tests and CAT indicated that the two-stage tests with the peaked routing tests measured as well as CATs at around mean ability level. Outside of these ability levels, they needed 30 to 40 percent more items to measure as well as the CATs at lower and higher end of ability levels. The relative efficiency of the two-stage tests with the rectangular routing tests was relatively consistent across the 16 ability levels. The average efficiency of the two-stage tests decreased with longer tests because the amount of information gain of the CAT exceeded that of two-stage tests with additional items. For example, the relative efficiency ratio of the 40-item two-stage tests with 8 second-stage tests were 89% (peaked routing test) or 88% (rectangular routing test) but these decreased into 85% (peaked) and 82% (rectangular) with the 45-item test and 81% (peaked) and 78% (rectangular) with the 50-item test.

It is evident from the results obtained that a fixed-length CAT is superior to the IRT-based two-stage tests of equivalent length in terms of measurement accuracy and efficiency. Ability

estimates from the two-stage test using an odd number (7) of second-stage tests with the 20-item rectangular routing test can compare favorably with true thetas and with ability estimates from CAT.

However, although the evidence from the computer simulations is encouraging, its generalization to real data is questionable; therefore the findings from any computer simulations need to be verified in live-testing. Also the specific results obtained are limited in their generalizability, since the characteristics of the tests and the item pool will have an effect upon the results obtained. In this study, a modified one-parameter model was used as only b parameters (with fixed a and c parameters) were used to estimate the ability parameter. According to van de Vijver (1986), IRT models with guessing parameter tend to underestimate lower abilities while the one-parameter (Rasch) model consistently tend to overestimate ability at lower levels. Thus, the IRT model chosen could have affected the ability estimation result. In addition, for IRT-based two-stage tests, the rectangular routing tests were designed to sequence items from easy to hard items. Further research needs to address how item difficulty ordering affect in ability estimation of two-stage tests with real data.

Additional research to examine the effects of the proportion of the overlapping items in second-stage tests and increasing odd number of second-stage tests (3, 5, 7) would be of interest to compare to the results from this study.

The administration features of the two-stage tests can

potentially influence the results. If two-stage tests are administered by computer, ability parameters will be quickly estimated; in addition routing error will be reduced by the automatic routing process by computer. On the other hand, if a two-stage test is administered by paper-and-pencil, scoring will be delayed, resulting in some time lapse to administer second-stage tests. The degree to which such delays would influence actual test performance should be considered.

In addition to this, test environment for CAT differs from a paper-and-pencil test environment in that with CAT an examinee typically cannot skip an item or return to an item to reconsider his/her choice. These features of paper-and-pencil test allow the examinee some control over the testing situation which is absent in a typical CAT. Therefore, this kind of psychological effect in the test environment needs to be considered in order to compare directly ability estimates from CAT and two-stage tests with paper-and-pencil administration.

In conclusion, IRT-based two-stage tests using rectangular distribution of item difficulties in the routing test and an odd number of second-stage tests produced more accurate theta estimates than did other two-stage test configurations studied. Further these ability estimates were close to those from the fixed-length CATs. Considering the limitations of CAT implementation and with the practical advantages of two-stage test administration, IRT-based two-stage test may be practical and feasible alternative for applications involving a wide range of student ability in school

settings.

**REFERENCES**

Adema, J.J. (1991). The construction of customized Two-Stage Tests. Journal of Educational Measurement, 27, 241-253.

Angoff, W.H., & Huddleston, E.M. (1958). The multi-level experiment: A study of a two-level test system for the College Board Scholastic Aptitude Test (Statistical Report SR-58-21). Princeton, NJ, Educational Testing Service.

Assessment Systems Corporation. (1988). User's manual for the MicroCAT Testing System, Version 3.0. St. Paul, MN: Author.

Betz, N.E., & Weiss, D.J. (1973). An empirical study of computer-administered two-stage ability testing (Research Report No. 73-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Method Program.

Betz, N.E., & Weiss, D.J. (1974). Simulation studies of two-stage testing (Research Report No. 74-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Method Program.

Cleary, T.A., Linn, R.L., & Rock, D.A. (1968a). An exploratory study of programmed tests. Educational and Psychological Measurement, 28, 345-360.

Cleary, T.A., Linn, R.L., & Rock, D.A. (1968b).

Reproduction of total test score through the use of sequential programmed tests. Journal of Educational Measurement, 5, 183-187.

Hulin, C.L., Lissak, R.I., & Drasgow, F. (1982). Recovery of Two- and three-parameter logistic item characteristic curves: A Monte Carlo study. Applied Psychological Measurement, 6(3), 249-260.

Linn, R.L., Rock, D.A., & Cleary, T.A. (1969). The development and evaluation of several programmed testing methods. Educational and Psychological Measurement, 29, 129-146.

Lord, F.M. (1971). A theoretical study of Two-Stage testing. Psychometrika, 36, 227-242.

Lord, F.M. (1980). Application of Item Response Theory to Practical Testing Problems. Hillsdale, NJ: Erlbaum.

Loyd, B.H. (1984). Efficiency and precision in two-stage adaptive testing. Paper presented at the Annual Meeting of the Eastern Educational Research Association, West Palm Beach.

Moreno, K.E., Wetzel, C.D., McBride, J.R., Weiss, D.J. (1984). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and computerized adaptive testing (CAT) subtests. Applied Psychological Measurement, 8, 155-163.

Olsen, J.B., Maynes, D.M., Slawson, D.A., & Ho, K. (1986). Comparison and equating of paper-administered,

computer-administered and computerized adaptive tests of achievement. Paper presented at the meeting of the American Educational Research Association, San Francisco.

van de Vijver, F.J.R. (1986). The robustness of Rasch estimates. Applied Psychological Measurement, 10,(1), 45-57.

Weiss, D.J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492.

Table 1

Correlations among Ability Estimates from 40-item Two-Stage Tests, those from CAT40, and True Ability

| True Θ | CAT40 | P-6 | P-7 | P-8 | R-6 | R-7 | R-8 |
|---|---|---|---|---|---|---|---|
| 1.000 | 0.983 | 0.977 | 0.971 | 0.976 | 0.976 | 0.975 | 0.976 |
| | 1.000 | 0.960 | 0.954 | 0.959 | 0.959 | 0.959 | 0.960 |

Correlations among Ability Estimates from 45-item Two-Stage Tests, those from CAT45, and True Ability

| True Θ | CAT45 | P-6 | P-7 | P-8 | R-6 | R-7 | R-8 |
|---|---|---|---|---|---|---|---|
| 1.000 | 0.985 | 0.979 | 0.976 | 0.978 | 0.979 | 0.981 | 0.979 |
| | 1.000 | 0.965 | 0.962 | 0.964 | 0.965 | 0.966 | 0.964 |

Correlations among Ability Estimates from 50-item Two-Stage Tests, those from CAT50, and True Ability

| True Θ | CAT50 | P-6 | P-7 | P-8 | R-6 | R-7 | R-8 |
|---|---|---|---|---|---|---|---|
| 1.000 | 0.987 | 0.982 | 0.978 | 0.980 | 0.982 | 0.982 | 0.981 |
| | 1.000 | 0.969 | 0.965 | 0.967 | 0.968 | 0.968 | 0.968 |

Note. Number (6, 7, 8) corresponds to the number of second-stage tests involved. P represents peaked routing test. R represents rectangular routing test.

Table 2

Average RMSE of Theta Estimates

| Test Length | | P-6 | P-7 | P-8 | R-6 | R-7 | R-8 | CAT |
|---|---|---|---|---|---|---|---|---|
| 40 | m | 0.404 | 0.420 | 0.402 | 0.406 | 0.411 | 0.405 | 0.350 |
| | sd | 0.074 | 0.116 | 0.075 | 0.054 | 0.058 | 0.050 | 0.026 |
| 45 | m | 0.381 | 0.392 | 0.385 | 0.379 | 0.364 | 0.379 | 0.322 |
| | sd | 0.081 | 0.132 | 0.099 | 0.033 | 0.026 | 0.048 | 0.023 |
| 50 | m | 0.357 | 0.370 | 0.369 | 0.359 | 0.356 | 0.360 | 0.314 |
| | sd | 0.065 | 0.127 | 0.087 | 0.026 | 0.022 | 0.031 | 0.031 |

Note. Number (6, 7, and 8) corresponds to the number of second-stage tests involved. P represents peaked routing test. R represents rectangular routing test. m is the mean RMSE. sd is the standard deviation of the mean RMSE.

Figure 1.1 MEAN RMSE

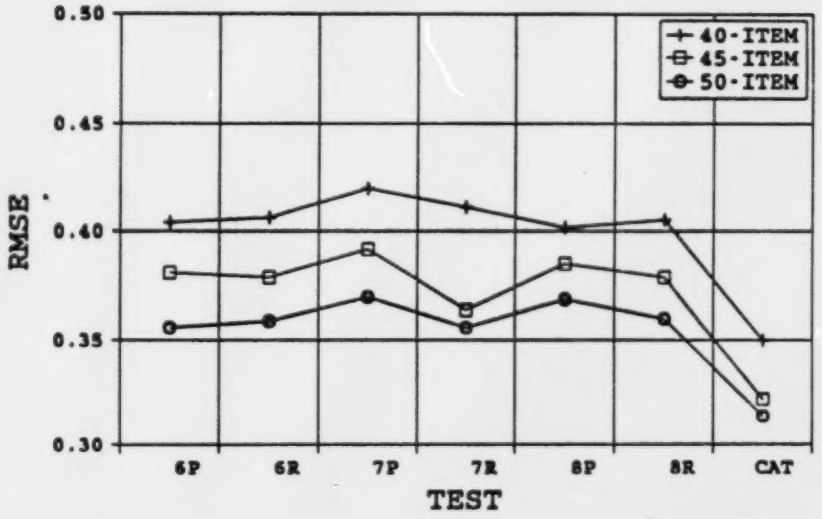



Figure 1.2 MEAN RMSE (PEAKED)

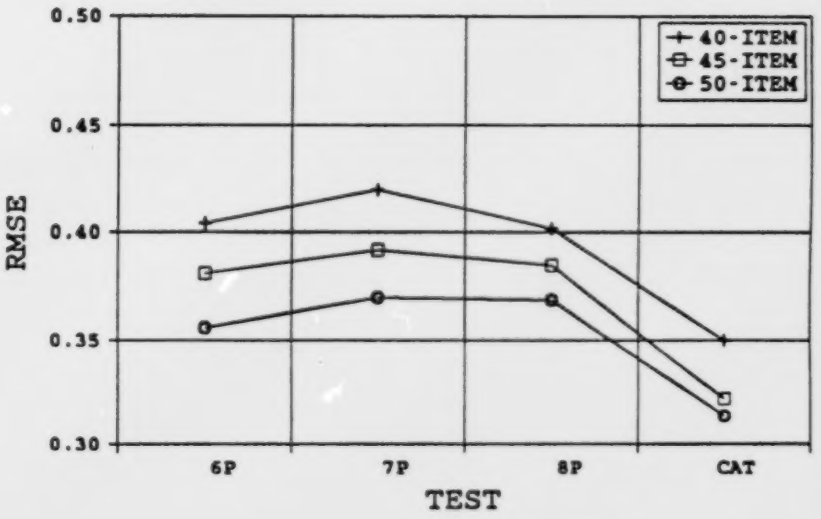
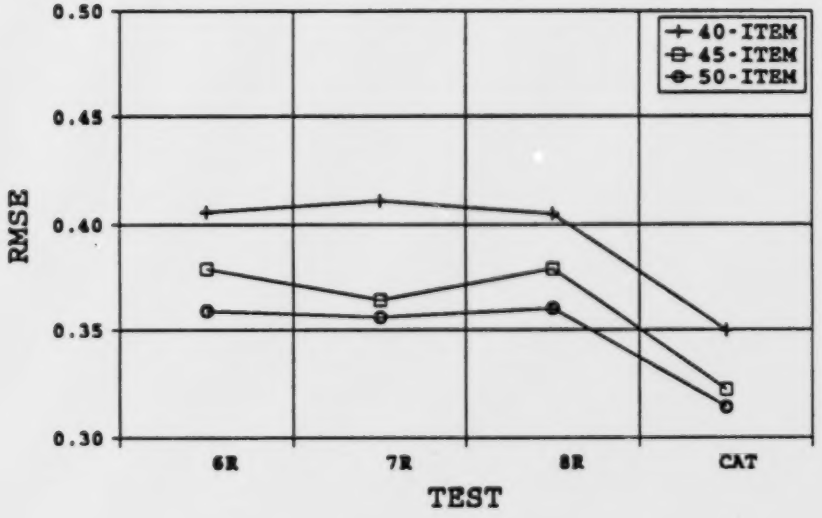



Figure 1.3 MEAN RMSE (RECTANGULAR)




Note: number (6, 7, and 8) corresponds to the number of second-stage tests involved. P represents peaked routing test. R represents rectangular routing test.

Figure 2.1 MEAN RMSE (PEAKED)
40-ITEM TESTS

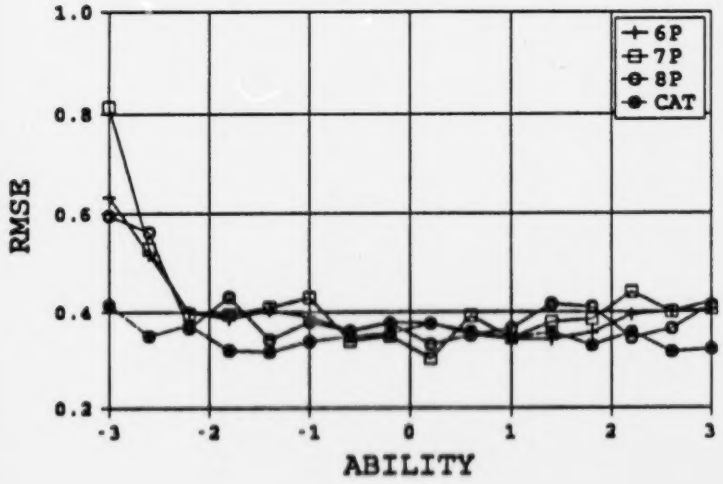



Figure 2.2 MEAN RMSE (PEAKED)
45-ITEM TESTS

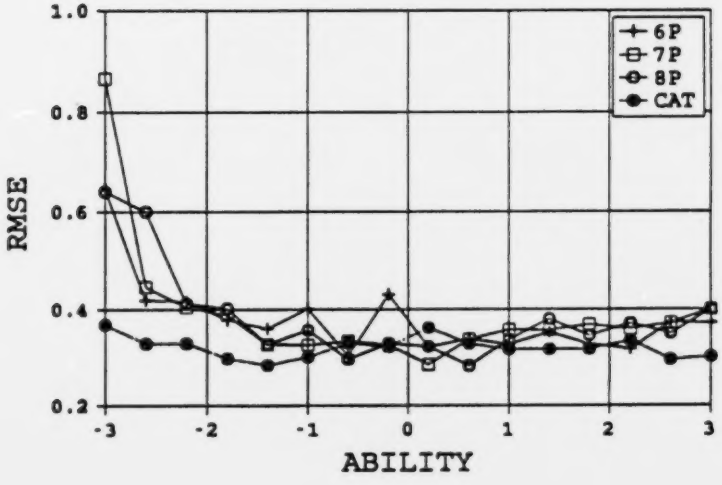
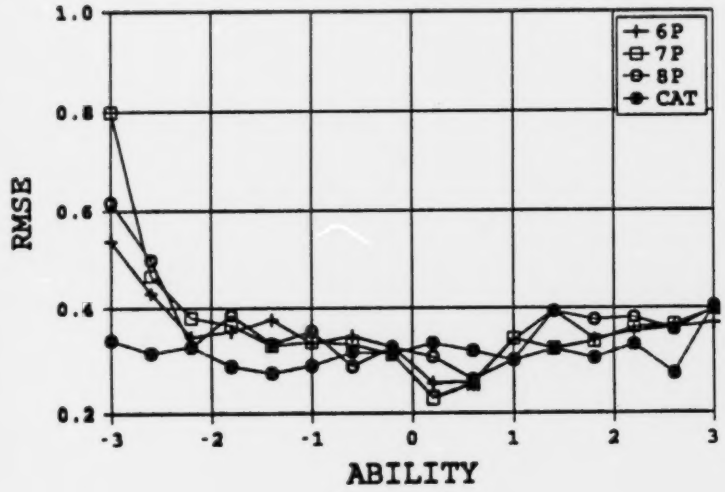



Figure 2.3 MEAN RMSE (PEAKED)
50-ITEM TESTS




Note: number (6, 7, and 8) corresponds to the number of second-stage tests involved. P represents peaked routing test.

Figure 3.1 MEAN RMSE (RECTANGULAR) 40-ITEM TESTS

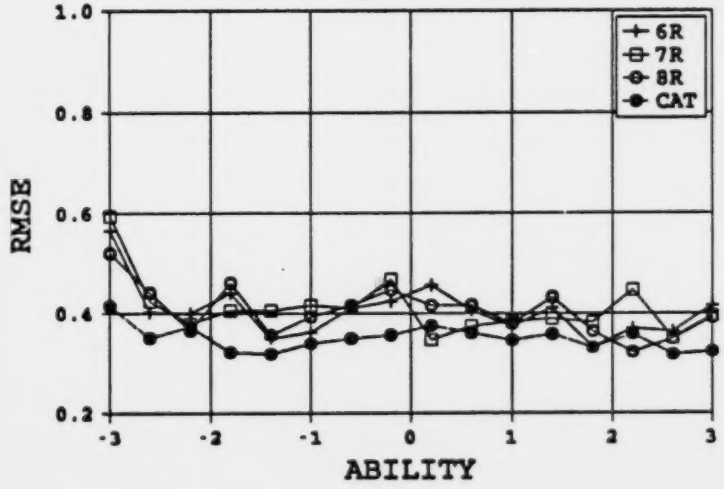



Figure 3.2 MEAN RMSE (RECTANGULAR) 45-ITEM TESTS

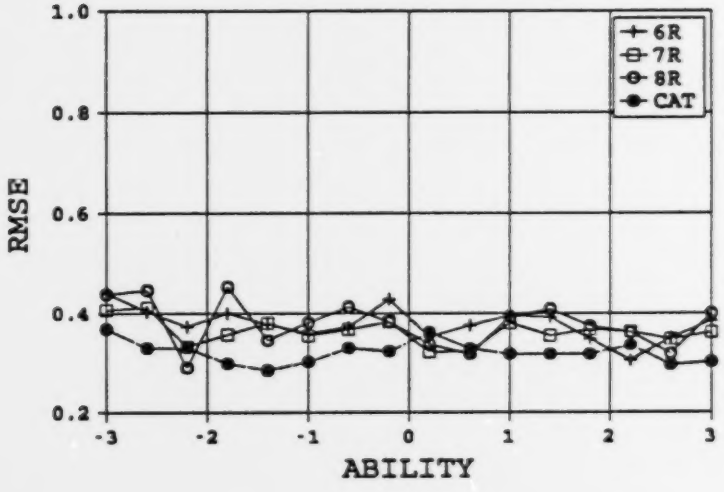
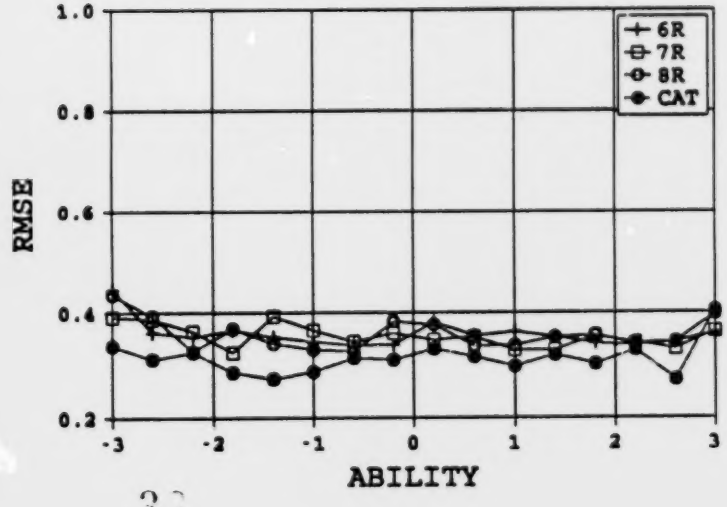



Figure 3.3 MEAN RMSE (RECTANGULAR) 50-ITEM TESTS




Note: number (6, 7, and 8) corresponds to the number of second-stage tests involved. R represents rectangular routing test.

Figure 4.1 MEAN BIAS (PEAKED)
40-ITEM TESTS

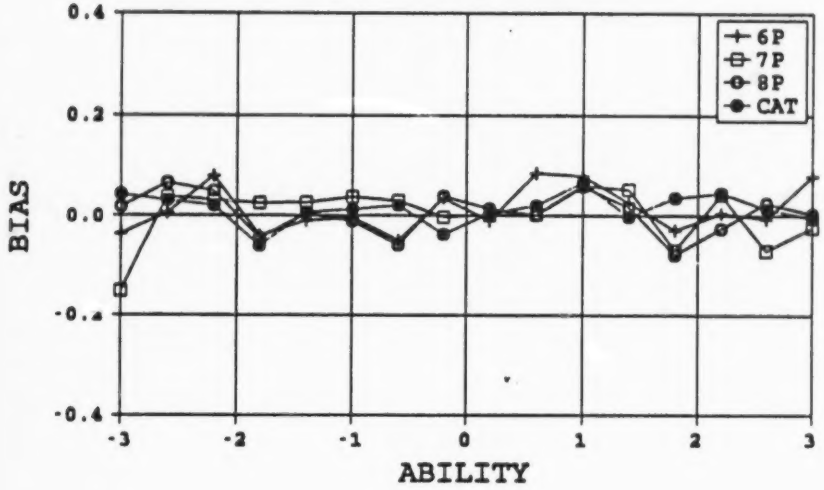



Figure 4.2 MEAN BIAS (PEAKED)
45-ITEM TESTS

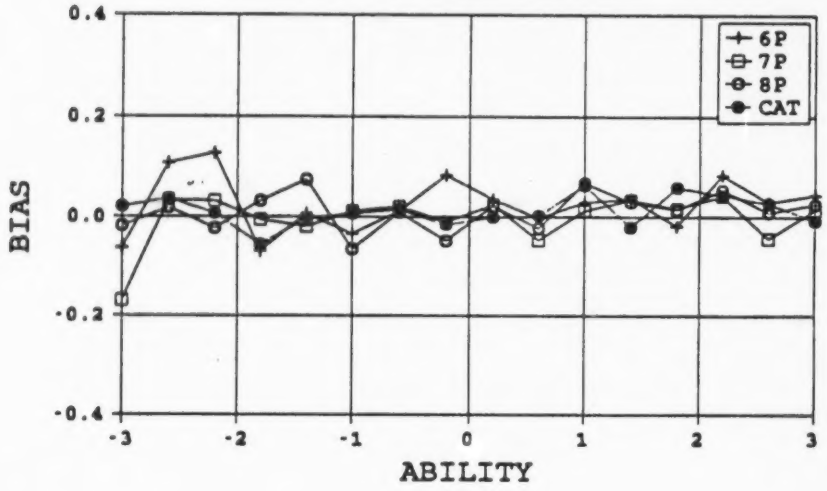
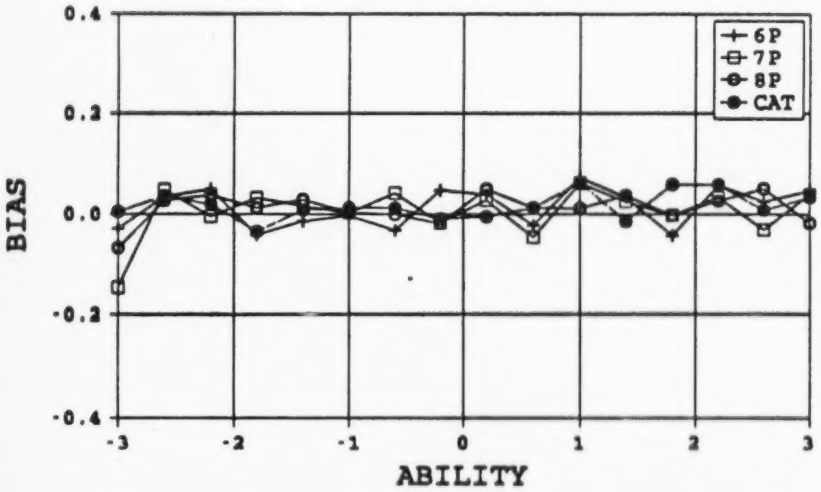



Figure 4.3 MEAN BIAS (PEAKED)
50-ITEM TESTS




Note: number (6, 7, and 8) corresponds to the number of second-stage tests involved. P represents peaked routing test.

Figure 5.1 MEAN BIAS (RECTANGULAR) 40-ITEM TESTS

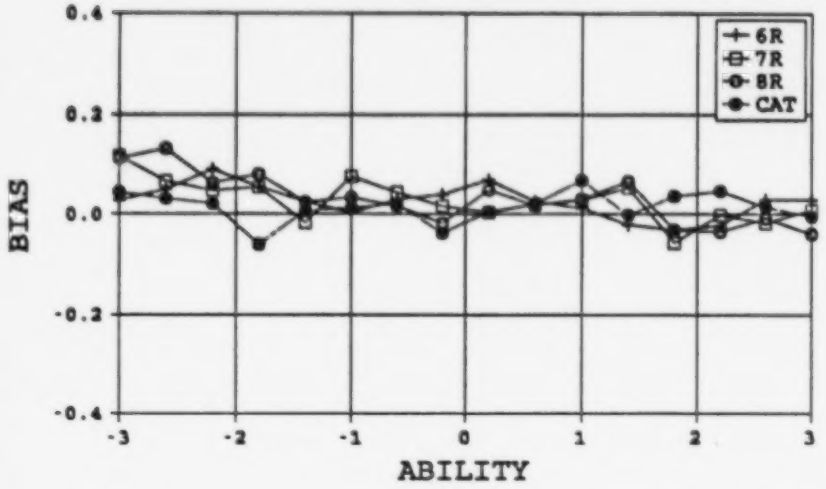



Figure 5.2 MEAN BIAS (RECTANGULAR) 45-ITEM TESTS

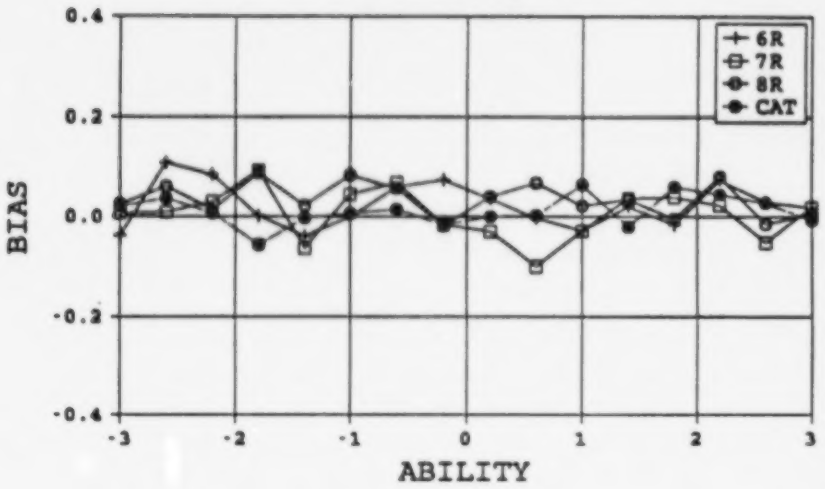
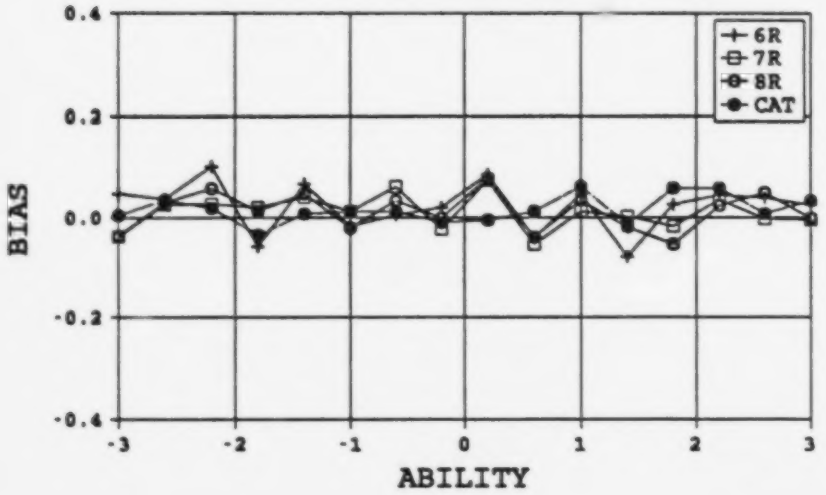



Figure 5.3 MEAN BIAS (RECTANGULAR) 50-ITEM TESTS




Note: number (6, 7, and 8) corresponds to the number of second-stage tests involved. R represents rectangular routing test.

Figure 6-1 Relative Efficiency of 40-Item Two-Stage Tests and CAT40

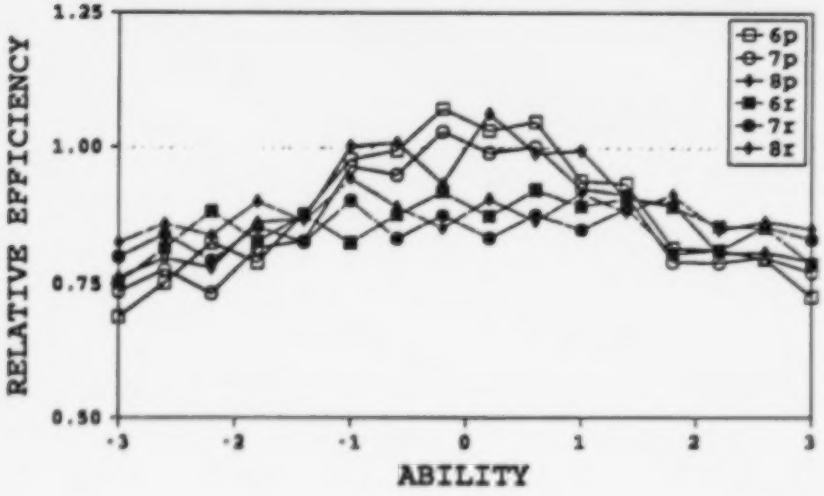



Figure 6-2 Relative Efficiency of 45-Item Two-Stage Tests and CAT45

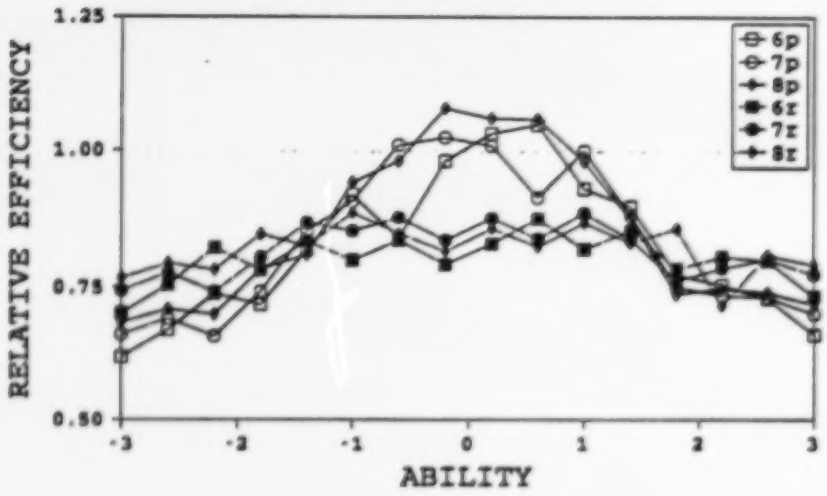
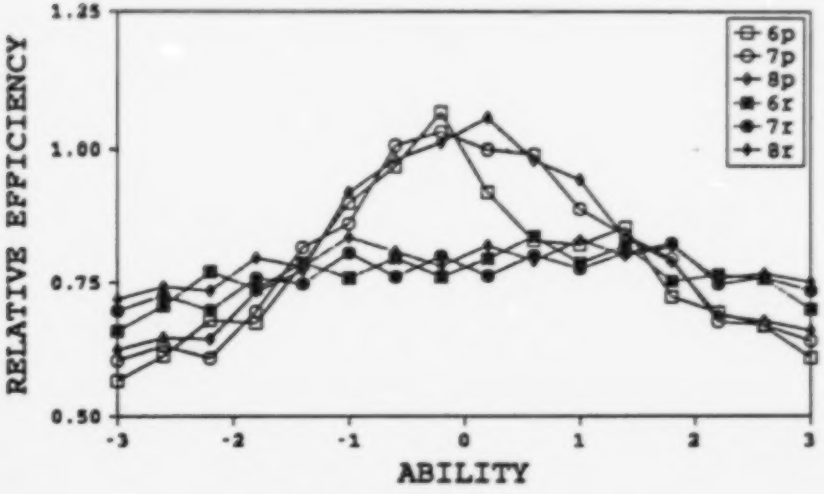



Figure 6-3 Relative Efficiency of 50-Item Two-Stage Tests and CAT50




Note: number (6, 7, and 8) corresponds to the number of second-stage tests involved. P represents peaked routing test. R represents rectangular routing test.